Report6

Load the R packages

```
library(knitr)
library(readr)
library(tidyverse)
library(dplyr)
library(tidyr)
library(stringr)
library(purrr)
library(tibble)
library(readxl)
```

Preparations: read the Excel data and do some quick renaming and check what happened at the 1956 games:

```
# Read data
athletes <- read_excel("olympics.xlsx", sheet="athletes")
games    <- read_excel("olympics.xlsx", sheet="games")
country  <- read_excel("olympics.xlsx", sheet="country")
medals   <- read_excel("olympics.xlsx", sheet="medals")
# Rename column for later JOINS
athletes <- athletes %>%
  rename(athlete_id = ID)
# What happened at the 1956 games?
games %>% filter(Year==1956)
```

```
## # A tibble: 3 x 4
##   Games        Year Season City
##   <chr>       <dbl> <chr>  <chr>
## 1 1956 Winter  1956 Winter Cortina d'Ampezzo
## 2 1956 Summer  1956 Summer Melbourne
## 3 1956 Summer  1956 Summer Stockholm
```

There were actually two Summer Games in 1956 - one in Melbourne and one in Stockholm!

Part 1

Have some athletes competed for different countries over time?

```
# Do the necessary JOINS
part1 <- athletes %>%
  full_join(country)
# Check summary statistics and remove duplicates
part1 <- part1 %>%
  group_by(Name, NOC) %>%
```

```
  select(Name, NOC) %>%
  distinct() %>%
  group_by(Name) %>%
  summarize(NUMBER_OF_COUNTRIES = n())
```

Indeed, 1748 athletes had various country affiliations.

Part 2

Who are the ten athletes that took part in most games?

```
# Do the necessary JOINS
part2 <- athletes %>%
  full_join(country)
# Summary statistics
part2 %>%
  group_by(Name) %>%
  summarize(NUMBER_OF_GAMES = n()) %>%
  arrange(desc(NUMBER_OF_GAMES)) %>%
  head(10)
```

```
## # A tibble: 10 x 2
##    Name                            NUMBER_OF_GAMES
##    <chr>                                     <int>
##  1 Ian Millar                                   10
##  2 Afanasijs Kuzmins                             9
##  3 Hubert Raudaschl                              9
##  4 Aleksandr Vladimirovich Popov                 8
##  5 Chen Jing                                     8
##  6 Durward Randolph Knowles                      8
##  7 Francisco Boza Dibos                          8
##  8 Josefa Idem-Guerrini                          8
##  9 Lesley Allison Thompson-Willie                8
## 10 Li Na                                         8
```

Part 3

What athlete(s) kept a Gold medal for the longest time?

```
# Do the necessary JOINS
part3 <- athletes %>%
  full_join(medals) %>%
  filter(Medal == 'Gold') %>%
  group_by(Name, Event) %>%
  mutate(NUMBER_OF_GOLD_BY_EVENT = n()) %>%
  ungroup() %>%
  filter(NUMBER_OF_GOLD_BY_EVENT>1) %>%
  arrange(desc(NUMBER_OF_GOLD_BY_EVENT)) %>%
  select(Name, Sex, Games, Team, Sport, Medal, NUMBER_OF_GOLD_BY_EVENT)
  head(10)
```

```
## [1] 10
```

```
# Print
part3
```

```
## # A tibble: 2,916 x 7
##    Name          Sex   Games    Team     Sport    Medal NUMBER_OF_GOLD_BY_~
##    <chr>         <chr> <chr>    <chr>    <chr>    <chr>               <int>
##  1 Aladr Gerevi~ M     1932 Su~ Hungary  Fencing  Gold                    6
##  2 Aladr Gerevi~ M     1936 Su~ Hungary  Fencing  Gold                    6
##  3 Aladr Gerevi~ M     1948 Su~ Hungary  Fencing  Gold                    6
##  4 Aladr Gerevi~ M     1952 Su~ Hungary  Fencing  Gold                    6
##  5 Aladr Gerevi~ M     1956 Su~ Hungary  Fencing  Gold                    6
##  6 Aladr Gerevi~ M     1960 Su~ Hungary  Fencing  Gold                    6
##  7 Reiner Klimke M     1964 Su~ Germany  Equestr~ Gold                    5
##  8 Reiner Klimke M     1968 Su~ West Ge~ Equestr~ Gold                    5
##  9 Reiner Klimke M     1976 Su~ West Ge~ Equestr~ Gold                    5
## 10 Reiner Klimke M     1984 Su~ West Ge~ Equestr~ Gold                    5
## # ... with 2,906 more rows
```

Considering Bobb'y comment "With keeping a gold medal we mean a gold medal on the same event", the tibble above shows that the Hungarian athlete *Aladr Gerevich (-Gerei)* won 6 gold medals in *Fencing Men's Sabre, Team* during at the following games: 1932 Summer, 1936 Summer, 1948 Summer, 1952 Summer, 1956 Summer, 1960 Summer.

Part 4

Based on the tibble of Part 3, Hungary kept a Gold medal for the longest time.

Part 5

Who are the ten athletes that competed in the most events (some athletes take part in more than one event during games) ?

```
# Do the necessary JOINS
part5 <- athletes %>%
  full_join(country) %>%
  full_join(medals) %>%
  distinct() %>%
  group_by(Name, Games) %>%
  summarise(NUMBER_OF_COMPETITIONS = n()) %>%
  arrange(desc(NUMBER_OF_COMPETITIONS)) %>%
  head(10)
# Print
part5
```

```
## # A tibble: 10 x 3
## # Groups:   Name [9]
##    Name                                Games      NUMBER_OF_COMPETITIO~
##    <chr>                               <chr>                      <int>
##  1 Willis Augustus Lee, Jr.            1920 Summ~                    15
```

```
##  2 Lloyd Spencer Spooner                    1920 Summ~                    13
##  3 Carl Schuhmann                           1896 Summ~                    12
##  4 Ioannis Theofilakis                      1912 Summ~                    12
##  5 Jean Fouconnier                          1906 Summ~                    12
##  6 "Marie Joseph \"Raoul\" le Borgne de B~ 1906 Summ~                    12
##  7 Maurice Faure                            1906 Summ~                    12
##  8 Bruno Julius Wagner                      1906 Summ~                    11
##  9 Frangiskos D. Mavrommatis                1912 Summ~                    11
## 10 Ioannis Theofilakis                      1920 Summ~                    11
```

Part 6

Create a new table showing the number of medals per country (rows) and per year (column). Keep only the 15 countries with the most medals overall.

```
# Compute the 15 countries with most medals
part6_15countries <- games %>%
  full_join(medals) %>%
  full_join(athletes) %>%
  select(Year, Team, Medal) %>%
  filter((Medal=='Gold') | (Medal=='Silver') | (Medal=='Bronze')) %>%
  group_by(Team) %>%
  summarize(TOTAL_MEDALS_OVERALL = n()) %>%
  arrange(desc(TOTAL_MEDALS_OVERALL)) %>%
  head(15)
# Print
part6_15countries
```

```
## # A tibble: 15 x 2
##    Team          TOTAL_MEDALS_OVERALL
##    <chr>                        <int>
##  1 United States                 5340
##  2 Soviet Union                  2620
##  3 Germany                       2036
##  4 Great Britain                 1710
##  5 France                        1583
##  6 Italy                         1572
##  7 Sweden                        1462
##  8 Australia                     1368
##  9 Canada                        1262
## 10 Hungary                       1191
## 11 Russia                        1110
## 12 Netherlands                    988
## 13 East Germany                   941
## 14 Japan                          935
## 15 Norway                         913
```

```
# Pull the country names for later comparison
top_countries <- pull(part6_15countries, Team)
# Now the same joins again but keep only the 15 countries with most medals
part6_medals <- games %>%
  full_join(medals) %>%
```

```
  full_join(athletes) %>%
  select(Year, Team, Medal) %>%
  filter(str_detect(Team, paste(top_countries, collapse = "|"))) %>%
  filter(str_detect(Team, "-", negate = TRUE)) %>%
  filter((Medal=='Gold') | (Medal=='Silver') | (Medal=='Bronze')) %>%
  group_by(Year, Team) %>%
  mutate(TOTAL_MEDALS = n()) %>%
  distinct() %>%
  ungroup() %>%
  arrange(Year) %>%
  select(Year, Team, TOTAL_MEDALS)
# Print a snippet
part6_medals %>% head(10)
```

```
## # A tibble: 10 x 3
##     Year Team                  TOTAL_MEDALS
##    <dbl> <chr>                        <int>
##  1  1896 Great Britain                    7
##  2  1896 United States                   20
##  3  1896 Germany                         31
##  4  1896 Great Britain                    7
##  5  1896 Great Britain/Germany            2
##  6  1896 United States                   20
##  7  1896 France                          11
##  8  1896 United States                   20
##  9  1896 Hungary                          6
## 10  1896 Great Britain                    7
```

Part 7

Create a scatterplot showing the average height and weight of competitors per sport (one dot per sport).

```
# Do the JOINS and compute the quantities
part7 <- athletes %>%
  full_join(medals) %>%
  group_by(Sport) %>%
  mutate(AVERAGE_HEIGHT = mean(Height, na.rm = TRUE)) %>%
  mutate(AVERAGE_WEIGHT = mean(Weight, na.rm = TRUE)) %>%
  mutate(AVERAGE_BMI    = mean((Weight / ((Height/100)*(Height/100))), na.rm = TRUE)) %>%
  select(AVERAGE_HEIGHT, AVERAGE_WEIGHT, AVERAGE_BMI, Sport) %>%
  ungroup() %>%
  distinct()
# Print some values
part7 %>% head(10)
```

```
## # A tibble: 10 x 4
##    AVERAGE_HEIGHT AVERAGE_WEIGHT AVERAGE_BMI Sport
##             <dbl>          <dbl>       <dbl> <chr>
##  1           191.           85.8        23.3 Basketball
##  2           174.           78.8        25.6 Judo
##  3           175.           70.4        22.8 Football
##  4           182.           95.6        27.5 Tug-Of-War
```

```
##  5          174.          70.0        23.0 Speed Skating
##  6          173.          65.9        21.8 Cross Country Skiing
##  7          176.          69.2        22.1 Athletics
##  8          179.          80.8        25.1 Ice Hockey
##  9          179.          70.6        22.0 Swimming
## 10          174.          68.2        22.4 Badminton
```

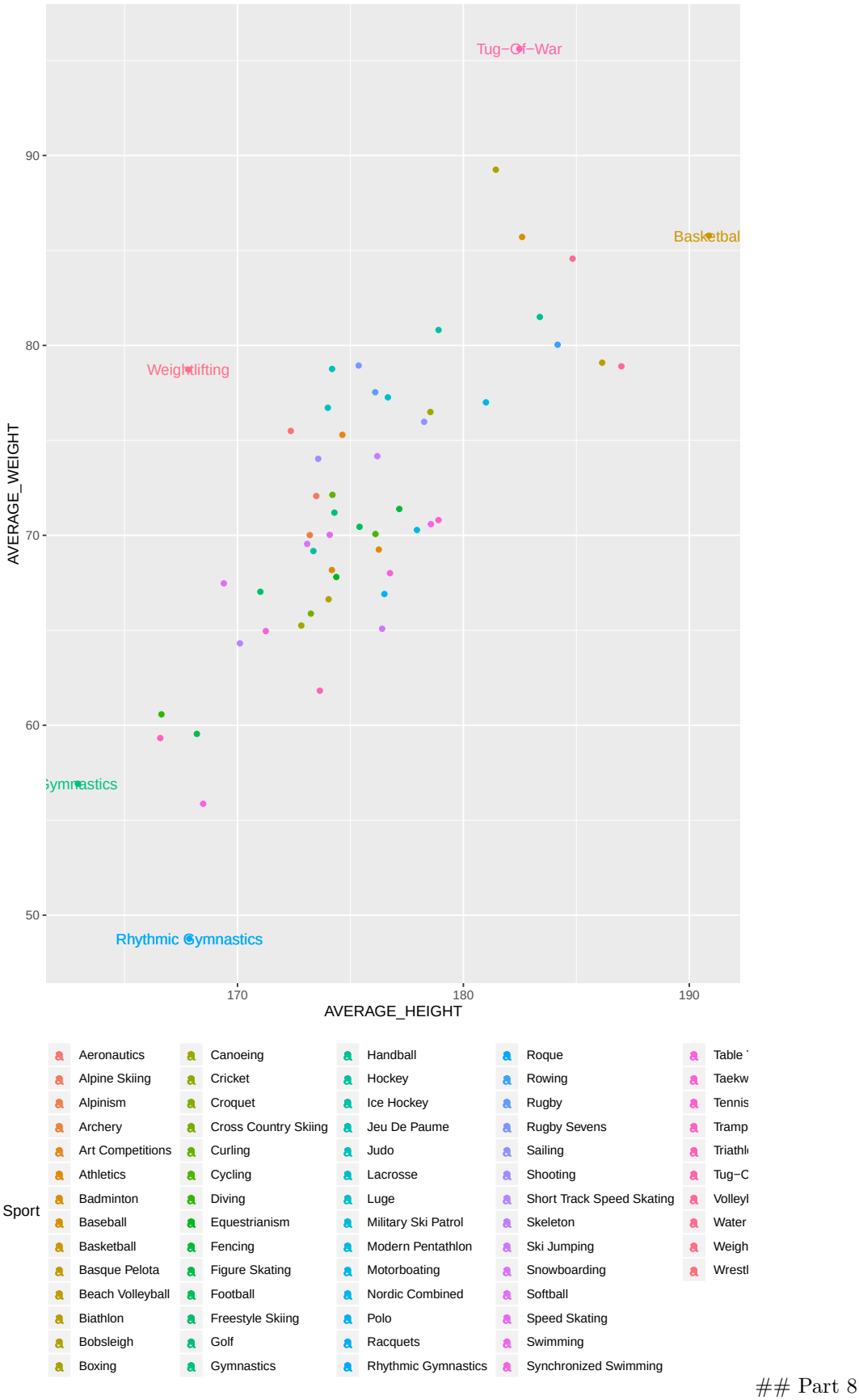
Computer other variables

```
# Include the other quantities (nested, not piped)
MAX_AVERAGE_HEIGHT = pull(arrange(part7,desc(AVERAGE_HEIGHT))["Sport"])[1]
MIN_AVERAGE_HEIGHT = pull(arrange(part7,(AVERAGE_HEIGHT))["Sport"])[1]
MAX_AVERAGE_WEIGHT = pull(arrange(part7,desc(AVERAGE_WEIGHT))["Sport"])[1]
MIN_AVERAGE_WEIGHT = pull(arrange(part7,(AVERAGE_WEIGHT))["Sport"])[1]
MAX_AVERAGE_BMI    = pull(arrange(part7,desc(AVERAGE_BMI))["Sport"])[1]
MIN_AVERAGE_BMI    = pull(arrange(part7,(AVERAGE_BMI))["Sport"])[1]
```

The sport with the largest average height is Basketball. The sport with the smaller average height is Gymnastics. The sport with the largest average weight is Tug-Of-War. The sport with the smaller average weight is Rhythmic Gymnastics. The sport with the largest average BMI is Weightlifting. The sport with the smaller average BMI is Rhythmic Gymnastics.

```
part7_filter <- filter(part7, Sport==MAX_AVERAGE_HEIGHT)

ggplot(part7, aes(x=AVERAGE_HEIGHT, y=AVERAGE_WEIGHT, color=Sport)) +
  geom_point() +
  theme(legend.position="bottom") +
  geom_text(data = filter(part7, Sport==MAX_AVERAGE_HEIGHT), aes(label = Sport)) +
  geom_text(data = filter(part7, Sport==MIN_AVERAGE_HEIGHT), aes(label = Sport)) +
  geom_text(data = filter(part7, Sport==MAX_AVERAGE_WEIGHT), aes(label = Sport)) +
  geom_text(data = filter(part7, Sport==MIN_AVERAGE_WEIGHT), aes(label = Sport)) +
  geom_text(data = filter(part7, Sport==MAX_AVERAGE_BMI), aes(label = Sport)) +
  geom_text(data = filter(part7, Sport==MIN_AVERAGE_BMI), aes(label = Sport))
```


Tug−Of−War
Basketbal
Weightlifting
Rhythmic Gymnastics
AVERAGE_WEIGHT
AVERAGE_HEIGHT
90
80
70
60
50
170
180
190
Sport
Aeronautics
Alpine Skiing
Alpinism
Archery
Art Competitions
Athletics
Badminton
Baseball
Basketball
Basque Pelota
Beach Volleyball
Biathlon
Bobsleigh
Boxing
Canoeing
Cricket
Croquet
Cross Country Skiing
Curling
Cycling
Diving
Equestrianism
Fencing
Figure Skating
Football
Freestyle Skiing
Golf
Gymnastics
Handball
Hockey
Ice Hockey
Jeu De Paume
Judo
Lacrosse
Luge
Military Ski Patrol
Modern Pentathlon
Motorboating
Nordic Combined
Polo
Racquets
Rhythmic Gymnastics
Roque
Rowing
Rugby
Rugby Sevens
Sailing
Shooting
Short Track Speed Skating
Skeleton
Ski Jumping
Snowboarding
Softball
Speed Skating
Swimming
Synchronized Swimming
Table
Taekw
Tennis
Tramp
Triathl
Tug−C
Volleyl
Water
Weigh
Wrestl

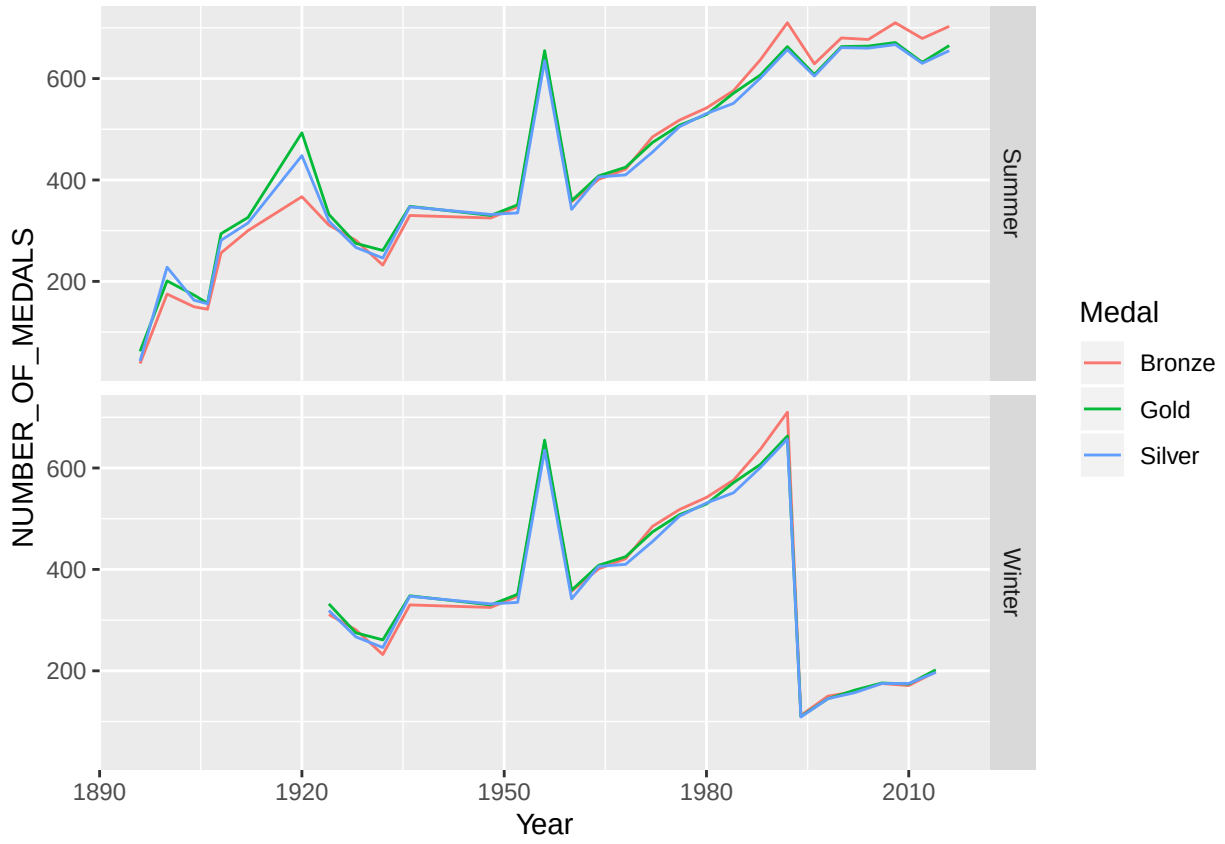
Part 8

Create a line plot showing the number of medals given by year (one line for Gold, one line for Silver and one line for Bronze). Does it change over time?

```
# Do the JOINS
part8 <- games %>%
  full_join(medals) %>%
  group_by(Year, Medal) %>%
  mutate(NUMBER_OF_MEDALS = n()) %>%
  ungroup() %>%
  distinct()
```

Do the plot:

```
# Do the JOINS
part8 %>%
  filter((Medal=='Gold') | (Medal=='Silver') | (Medal=='Bronze')) %>%
  ggplot(aes(x=Year, y=NUMBER_OF_MEDALS, colour=Medal)) +
  geom_line() +
  facet_grid(rows = vars(Season))
```



There is actually quite a significant change in the number of medals awarded over the years and in terms of Season (Summer vs. Winter).